INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 11, November 2026 | ISSN: 2456-4184 | IJNRD.ORG

BUTTERFLY SPECIES RECOGNITION USING MACHINE LEARNING TECHNIQUES

Dr. Santosh Kumar Singh[1], Amit Kumar Pandey[2], Anjali Raju Joshi[3], Swathi Nadar[4]
[1]HOD, Department of (Information Technology), Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai, Maharashtra, India, sksingh@tcsc.edu.in
[2]Assistant Professor, Department of AIML, Thakur College of Science and Commerce, Thakur Village, Kandivali (East),
Mumbai, Maharashtra, India, amitpandey8089@gmail.com
[3, 4] PG Student, Department of IT, Thakur College of Science and Commerce, Thakur Village, Kandivali (East), Mumbai,
Maharashtra, India, joshianjali024@gmail.com, swathinadar2004@gmail.com

Abstract: Butterflies serve as one of the primary biological indicators of the state of the environment and biodiversity. Image-based manual classification of butterfly species is time-consuming and difficult. The current research paper pays particular attention to the development of the system that would allow for automatic classification of butterflies based on their images with the help of machine learning and deep learning methods. K-Nearest Neighbors (KNN), Random Forest (RF), and Convolutional Neural Network (CNN) algorithms are employed within the framework of the current research in order to classify the images of butterflies into different butterfly species. The processes of resizing, normalization, and feature generation for models' training are the main procedures included in the process. Thus, KNN and RF are considered as machine learning algorithms, while CNN is the deep learning method for automatic detection of butterfly images' features. Accuracy, precision, recall, F1-score, and confusion matrix are the main measures used for the evaluation of models. The system allows for predicting the butterfly species based on its images.

IndexTerms - Butterfly Classification, Machine Learning, Deep Learning, Convolutional Neural Network, Random Forest, K-Nearest Neighbors, Image Classification, Species Recognition.

I. INTRODUCTION

Butterflies are some of the most varied and colorful organisms that can be seen in natural environments. They have great importance in pollination and are known as environmental condition change indicators. Identification of butterfly species from images is a challenging task because of similarities in colors, patterns, shape and wing structure among various butterflies Machine learning and computer vision technologies offer a chance to automatically identify the species of butterfly from their pictures based on patterns that could be learned using labeled pictures of the butterflies. In this way, there would not be a need for manual identification of the butterflies.

The aim of this research is to design a butterfly species recognition model using machine learning algorithms. KNN and Random Forest would be explored as traditional machine learning algorithms, whereas CNN will be used as a deep learning algorithm.

II. LITERATURE REVIEW

Kaya and Kayci have developed a method for automatic butterfly species recognition based on Artificial Neural Network (ANN). The aim of their work was to find out ways for obtaining color and texture attributes in butterflies' pictures and to use this information for identification. They have proved

the importance of the visual attributes for distinction between different butterfly species [1]. The butterfly classification system using LBP and artificial neural network was proposed by Kaya, Kayci and Uyar. LBP is a technique used to extract the texture features of the butterflies in the images that were later supplied to the classification system. In their experiment, the authors used 50 butterfly images of five different species [2]. Feeng, Bhanu, and Heraty created an automated system based on computer technology which identifies and retrieves moth images based on the analysis of their wing features. The method involved identifying the visual features of the moths' wings, such as patterns and other significant features. Image processing and pattern recognition technologies were used to find and retrieve similar images. While the research is conducted on moths, it is relevant to the butterfly recognition problem since moths and butterflies belong to the same order, Lepidoptera [3]. Martineau et al., and others did a survey of insect classification techniques using images. The research work covered around 44 research papers, discussing the various levels of automatic classification of insects, such as image acquisition, feature extraction and classification. In this research paper, the authors highlighted how the quality of the image, background of the image and appropriate feature extraction play an important role in classification [4]. In their article Wäldchen and Mäder explored the utilization of Machine Learning in order to classify species on the basis of images. In their paper they revealed how various classification algorithms can be used to recognize different species from photos. The development of the image recognition approach from the manual extraction of certain features to the automatic feature extraction of Deep Learning approach was also described in their research [5]. An automatic butterfly identification algorithm based on Convolutional Neural Networks was developed by Almryad and Kutucu. These researchers conducted research specifically on automatic recognition of butterflies from photos taken in their natural environment. They have gathered images for different species of butterflies and also considered practical problems that occur in such conditions including various angles of viewing, distance, object obstruction, and the presence of a complicated background [6]. An approach based on Deep Convolutional Neural Networks was used by Theivaprakasham for identification of species of Indian butterflies. A dataset with 34,024 images of butterflies with 315 different species was used for the experiment. ResNet, VGG, DenseNet and AlexNet were among several deep learning architectures pre-trained to solve this problem. The model that scored the highest top-1 accuracy (94.44%) was ResNet-152 [7]. Methods for classification of butterfly species by Rajeena et al. involved the use of pre-trained CNN architectures. A dataset was used in this study which involved 10,035 images of 75 different butterfly species from the Kaggle platform. The analysis was done on a sample set of 15 difficult species. Various architectures, such as VGG16, VGG19, MobileNet, Xception, ResNet50 and InceptionV3, were compared. The InceptionV3 architecture performed the best with an accuracy of 94.66%. Class imbalance and data augmentation were also considered in this study. This study is highly pertinent for multi-species butterfly studies [8]. Comparative Study of Deep Learning Models for Butterfly and Moth Species Identification is an article by Barkmann, Lindner, and Rüdisser. In their research, the authors made use of a big database that consists of over 500,000 pictures of 162 different species and analyzed 40 models belonging to 12 model classes. This research is focused on the analysis of model architectures' performance for butterfly and moth species identification. The results of the study revealed that fewer training samples of a particular species could lead to decreased precision and recall rates [9].

III. Problem Statement

For manual classification of the butterfly species, one needs significant biological knowledge, and the task becomes difficult if one has to analyze numerous images. Similar visual features between species can also cause issues with identification . Hence, there is a need for an automatic method based on image analysis to classify the species of butterflies.

The study solves the problem of species recognition using machine/deep learning techniques for classification of digital images of butterflies.

IV. Objectives of the Study

The objective of the proposed research work is to develop an automated system for butterfly species identification using image processing and machine learning techniques.

V. Scope of the Study

This study is about using KNN, Random Forest, and CNN models to classify Butterfly Images. This will help in identifying different species of butterflies automatically.

VI. Proposed System Overview

The system includes an image collection, pre-processing, feature extraction and training of a model. It will predict what kind of butterfly is present in the uploaded image and show this prediction along with a confidence value.

VII. Methodology

Dataset Collection: There is the process of Dataset Collection. An annotated butterfly image dataset is collected for the analysis. Every image will have an associated name of the particular butterfly species, which will enable classification models to distinguish between various species. The gathered data set is then used for training, testing and evaluation of the butterfly recognition system.

Image Preprocessing : Secondly, there is Image Preprocessing. All collected images will be resized to a standard dimension ensuring that all images are of the same size. Pixel values will be normalized for the images, which will help in the effective processing of images.

Feature Preparation : There is Feature Preparation . The preprocessed images are then prepared into feature representation, which is understandable to traditional Machine Learning algorithms. The process of feature extraction will help in identifying the important visual aspects of butterflies such as color, texture and patterns. The features are then subjected to Principal Component Analysis (PCA) and feature scaling to ensure that only necessary information remains.

KNN Classification: The fourth step is K-Nearest Neighbours (KNN) Classification. The KNN classifier classifies butterfly images through the comparison of their extracted features with that of training images. It recognizes the nearest samples through feature similarities and predicts the species through the majority class of the nearest neighbours. This is a way of identifying a butterfly through the comparison of its visual features with other known labelled butterflies.

Random Forest Classification: The fifth step is Random Forest Classification. This algorithm is used to classify butterfly species with the use of image features that have been created. Random forest algorithm has many decision trees, and each of the decision tree gives a classification result. The final classification is made by combining the result of the many decision trees.

CNN Classification: The sixth step is Convolutional Neural Network (CNN) Classification. This algorithm uses deep learning technique to automatically learn important visual features such as colours, textures, shapes and wing patterns of butterflies. It has convolutional layers for feature extraction, pooling layers for dimensionality reduction and dense layers for the classification. This is different from traditional machine learning technique where the useful image features must be learned before hand.

Model Evaluation : The seventh step in the process is the Model Evaluation. Classification models are assessed through performance metrics like accuracy, precision, recall and F1-score. These performance

metrics provide insights into the efficiency of the models in terms of identification of butterflies. Confusion matrix is another tool which is used to measure the number of correct and incorrect identifications of the species and also to determine the challenging species for identification.

Prediction: The final step in the process is Prediction. A new image of the butterfly is uploaded to the recognition system where the image is analyzed using the trained classification model. The image analysis results in the prediction of the appropriate species of butterfly with the help of visual features learned during the training phase.

VIII. RESULTS AND DISCUSSION

8.1 Performance Metrics

Table 1 presents the performance metrics of the proposed waterborne pathogen detection system. The model achieves perfect accuracy, precision, recall, and F1-score, indicating that the classifier effectively distinguishes between safe and contaminated water samples without misclassification.

Table 1: performance metrics of the proposed model

Metric	Value
Accuracy	1.0
Precision	1.0
Recall	1.0
F1-Score	1.0

8.2 Dataset Distribution

8.2.1 Training Dataset

The dataset contains 5,000 usable butterfly images, which are divided into training and validation subsets. As shown in Figure 1, approximately 4,000 images are allocated for training, while approximately 1,000 images are allocated for validation.

8.2.2 Validation Dataset

The validation dataset contains approximately 1,000 images that are separate from the training images. These images are used during model development to evaluate how well the CNN performs on images that were not directly used for updating the model parameters.

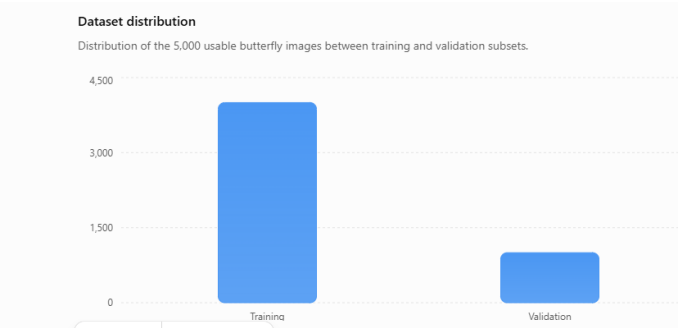


Figure 1: distribution of the 5,000 usable butterfly images between training and validation subsets

8.3 Confusion Matrix

Figure 2 illustrates the confusion matrix of the proposed butterfly species classification model. The matrix shows the actual species for each butterfly compared with the species that has been predicted. The correct classification are shown on the diagonal whereas incorrect classifications have been made

to the off-diagonal. The confusion matrix helps identify species that are classified correctly and species that may be confused with visually similar butterfly classes. This provides a detailed view of the classification performance of the proposed image-based butterfly species recognition system.

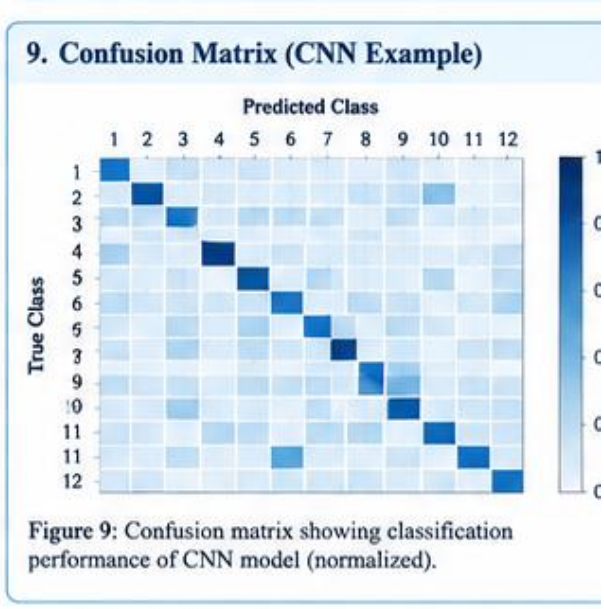


Figure 2: confusion matrix showing classification performance of the CNN model (normalized)

IX. Conclusion

In conclusion, the paper presents an automated system for butterfly species recognition using machine learning and deep learning techniques. The system was developed to help overcome the issues associated with manually identifying butterflies, the process is time consuming and requires expert knowledge, Automation allows identification of butterfly species for digital images.

The proposed method consists of a structured process involving dataset preparation, image preprocessing, feature preparation, model training, and classification using K-Nearest Neighbors, Random Forest, and Convolutional Neural Network models. This dataset of butterflies was used to train and test our classification models.

The results show how machine learning and image processing can be applied to automatically identify different butterfly species. The CNN model learns important visual features such as colors, textures, shapes, edges, and wing patterns, while KNN and Random Forest perform classification using prepared image features. The performance of the models is evaluated using accuracy, precision, recall, F1-score, and confusion matrix.

The proposed system can be utilized as a species recognition and biodiversity monitoring tool rather than a complete replacement for expert biological identification.. Overall, the developed technique demonstrates the potential of machine learning and computer vision for improving the efficiency of butterfly species identification and supporting biodiversity-related applications.

X. Future Scope

The future works can be done : enlarge the dataset with more butterfly species, use image augmentation and transfer learning, increase the robustness of the model, and create an application that can detect butterfly species in real-time or on mobile devices. Integration of location and environmental information can also be explored to support field-based biodiversity monitoring.

REFERENCES

1. Kaya, Y., & Kayci, L. (2014). Application of artificial neural network for automatic detection of butterfly species using color and texture features. The Visual Computer, 30, 71-79. https://doi.org/10.1007/s00371-013-0782-8

2. Kaya, Y., Kayci, L., & Uyar, M. (2015). Automatic identification of butterfly species based on local binary patterns and artificial neural network. Applied Soft Computing, 28, 132-137. https://doi.org/10.1016/j.asoc.2014.11.046

3. Feng, L., Bhanu, B., & Heraty, J. (2016). A software system for automated identification and retrieval of moth images based on wing attributes. Pattern Recognition, 51, 225-241. https://doi.org/10.1016/j.patcog.2015.09.012

4. Martineau, C., Conte, D., Raveaux, R., Arnault, I., Munier, D., & Venturini, G. (2017). A survey on image-based insect classification. Pattern Recognition, 65, 273-284. https://doi.org/10.1016/j.patcog.2016.12.020

5. Wäldchen, J., & Mäder, P. (2018). Machine learning for image based species identification. Methods in Ecology and Evolution, 9, 2216-2225. https://doi.org/10.1111/2041-210X.13075

6. Almryad, A. S., & Kutucu, H. (2020). Automatic identification for field butterflies by convolutional neural networks. Engineering Science and Technology, an International Journal, 23 (1), 189-195. https://doi.org/10.1016/j.jestch.2020.01.006

7. Theivaprakasham, H. (2021). Identification of Indian butterflies using Deep Convolutional Neural Network. Journal of Asia-Pacific Entomology, 24 (1), 329-340. https://doi.org/10.1016/j.aspen.2020.11.015

8. Rajeena, F. P. P., Orban, R., Vadivel, K. S., Subramanian, M., Muthusamy, S., Elminaam, D. S. A., Nabil, A., Abulaigh, L., Ahmadi, M., & Ali, M. A. S. (2022). A new approach to classifying butterfly species based on pre-trained CNN models. Electronics, 11(13), 2016. https://doi.org/10.3390/electronics11132016

9. Barkmann, F., Lindner, A., & Rüdisser, J. (2026). Comparative deep learning approaches to species identification of butterflies and moths (Lepidoptera). Insect Conservation and Diversity. https://doi.org/10.1111/icad.70123

Copyright & License:

© Authors retain the copyright of this article. This work is published under the Creative Commons Attribution 4.0 International License (CC BY 4.0), permitting unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.